A Proposal of a Privacy-preserving Questionnaire by Non-deterministic Information and Its Analysis

A Proposal of a Privacy-preserving Questionnaire by Non-deterministic Information and Its Analysis

Hiroshi Sakai, Chenxi Liu
Department of Applied Science,
Graduate School of Engineering,
Kyushu Institute of Technology,
Tobata, Kitakyushu 804-8550, Japan
E-mail: sakai@mns.kyutech.ac.jp

Michinori Nakata
Faculty of Management and
Information Science,
Josai International University,
Togane, Chiba 283-8555, Japan
E-mail: nakatam@ieee.org

Shusaku Tsumoto
Department of Medical Informatics,
School of Medicine, Shimane University,
Enya-cho, Izumo,
Shimane 693-8501, Japan
E-mail: tsumoto@med.shimane-u.ac.jp

***Abstract*—We focus on a questionnaire consisting of three-choice question or multiple-choice question, and propose a privacy-preserving questionnaire by non-deterministic information. Each respondent usually answers one choice from the multiple choices, and each choice is stored as a tuple in a table data. The organizer of this questionnaire analyzes the table data set, and obtains rules and the tendency. If this table data set contains personal information, the organizer needs to employ the analytical procedures with the privacy-preserving functionality.**

In this paper, we propose a new framework that each respondent intentionally answers non-deterministic information instead of deterministic information. For example, he answers 'either A, B, or C' instead of the actual choice A, and he intentionally dilutes his choice. This may be the similar concept on the k-anonymity. Non-deterministic information will be desirable for preserving each respondent's information.

We follow the framework of *Rough Non-deterministic Information Analysis* (RNIA), and apply RNIA to the privacy-preserving questionnaire by non-deterministic information. In the current data mining algorithms, the tuples with non-deterministic information may be removed based on the data cleaning process. However, RNIA can handle such tuples as well as the tuples with deterministic information. By using RNIA, we can consider new types of privacy-preserving questionnaire.

I. Introduction

We are coping with rough set based information incompleteness, missing values, and data mining [7], [10], [11], [12], [14], [15], [16], [18], [19], [26], [27] in table data sets, and we propose a framework of a privacy-preserving questionnaire in this paper. The idea is simple, namely, each respondent may answer non-deterministic information [14] instead of deterministic information, like 'either A, B, or C' instead of the actual one value. Each respondent can dilute his actual answer, and he can preserve his personal information. Of course, such questionnaire will be more privacy-preserved, and we analyze such questionnaire based on *Rough Non-deterministic Information Analysis* (RNIA). Figure 1 shows the total chart for a standard questionnaire, and Figure 2 shows the total chart for the proposing questionnaire.

Recently, the privacy issue on data engineering is very important, and this is often dealt as privacy-preserving data mining [1], [2], [5], [6], [9], [13], [22]. In [1], [2], several approaches, for example, randomization, k-anonymization, distributed privacy-preserving data mining, etc. are summarized.

In the randomization method, the noise is added to data for masking the attribute values [5]. This randomization seems to be closely related to the proposing questionnaire. In [5], the

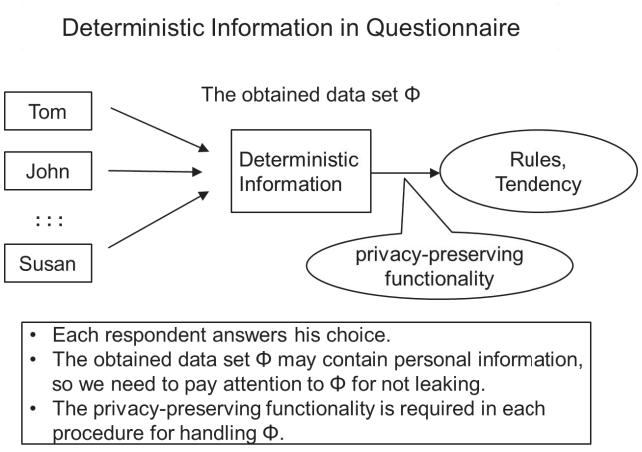


Fig. 1. The total chart handling data set Φ with deterministic information.

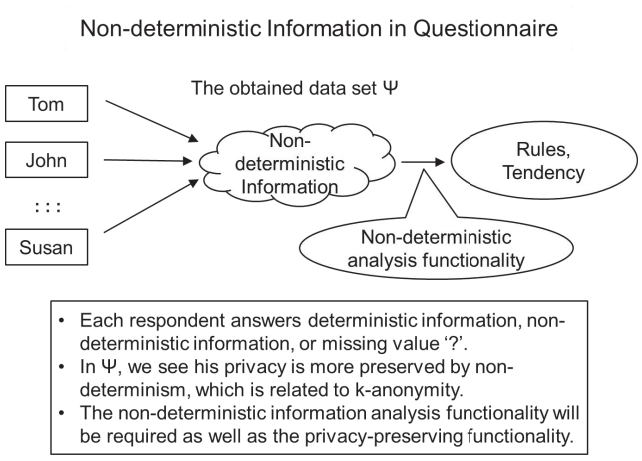


Fig. 2. The total chart handling data set Ψ with non-deterministic information.

organizer of the questionnaire adds noise to data, but each respondent adds noise to data in our questionnaire. Since the obtained data set by the proposing questionnaire stores vague information, we may have the weakened tendency. However, this will follow the description that the purpose of data mining is to obtain the general tendency of all respondents, and it is not to obtain each personal information [9].

The k-anonymity model was developed for not to identify any individual records [1], [2]. In our proposal, each respondent may answer 'either A, B, or C', which will be corresponding to 3-anonymity in an answer. This will be convenient for each respondent, because we often have information leaks.

In [6], [13], a privacy-preserving web-based questionnaire is investigated, and the issue on a secure protocol for handling the distributed data sets is considered. However, this is the different framework from our proposal.

In our proposal, the main issue is data mining in tables with non-deterministic information, so our proposal will be different from the traditional research on privacy-preserving data mining. However, we think data mining in tables with non-deterministic information is another approach to privacy-preserving.

This paper is organized as follows. In Section 2, we will propose a privacy-preserving questionnaire, and define a questionnaire *QUEST_Det* and a questionnaire *QUEST_Non-Det*. In Section 3, we consider NIS-Apriori based rule generation and a prototype system in SQL for handling *QUEST_Non-Det*. In Section 4, we discuss the merit and the demerit for *QUEST_Det* and *QUEST_Non-Det*. In Section 5, we apply the prototype in SQL to Mammographic data set and Lenses data set [8]. In Section 6, we show the use of NIS-Apriori based rule generation by *getRNIA* software tool opened in the web [28]. Finally, we conclude the possibility of applying non-deterministic information to a privacy-preserving questionnaire, and clarify the next research.

II. A Proposal of a Privacy-preserving Questionnaire

Let us consider the following questionnaire *QUEST_Det* by the organizer.

1) The questionnaire consists of some questions, which are multiple choices.
2) Each respondent answers one choice (deterministic information) from the multiple choices in each question.
3) The organizer analyzes the table data set, and obtains the rules for knowing the tendency in the respondents.

A. A Case of Deterministic Information in a Questionnaire

In this subsection, we consider three-choice questions for simplicity. We often have a questionnaire *QUEST_Det*, and the organizer may have a table data set Φ in Table I. We may call a table data set with definite information a *Deterministic Information System* (DIS) [14], [15], [16], [18], [19]. The organizer will have the public opinion and the tendency by analyzing Φ.

TABLE I
The obtained data set Φ based on three-choice questions.

Respondent	$q1$	$q2$	$q3$	$q4$
r_1	1	1	1	1
r_2	1	2	1	1
r_3	2	2	2	2
r_4	1	2	2	3
r_5	1	3	2	3
r_6	2	3	3	3

TABLE II
The obtained data sets Ψ.

Respondent	$q1$	$q2$	$q3$	$q4$
r_1	1	?	1	1
r_2	1	2	1	1
r_3	2	$\{1,2\}$	2	2
r_4	1	2	2	3
r_5	$\{1,2\}$	$\{1,2,3\}$	2	$\{2,3\}$
r_6	2	3	3	3

Related to the obtained data set Φ, we need to pay attention to the following.

1) If Φ contains personal and privacy information, it is necessary to assure the security of Φ. The privacy-preserving functionality is required in each procedure for handling Φ.
2) Usually, the security of the obtained data set will be managed by the organizer, however there have been the frequent leaks of data sets with personal information.
3) For each respondent, the most convenient answer for preserving his privacy may be the choice of either 'no answer' or 'either A, B, or C' (non-deterministic information) instead of the actual choice 1.

B. A Case of Non-deterministic Information in a Questionnaire

In order to preserve his personal information intentionally, we employ non-deterministic information instead of deterministic information. We define each non-deterministic information as a set S of choices, and we interpret S as that either an element of S is the actual choice but we do not know it. For example, non-deterministic information $\{1,2\}$ means the actual choice is 'either 1 or 2'. Non-deterministic information will be the similar concept of the k-anonymity, and it does not give definite information.

If the organizer of the questionnaire agrees with the use of non-deterministic information, each respondent can intentionally preserve his privacy in the question. Let us suppose we have table Ψ in Table II. In this case, the answers to $q2$ by r_1 and r_5 are semantically the same, because we identify the response 'no answer' expressed by '?' symbol with $\{1,2,3\}$ in the three-choice question. We may call a table data set with non-deterministic information a *Non-deterministic Information System* (NIS) [14], [15], [16], [18], [19].

Similarly to Φ, we need to pay attention to the following about Ψ.

1) If we have Ψ instead of Φ, personal information will be more preserved.

2) However, if we employ the traditional data mining algorithm, the tuples r_1, r_3, and r_5 may be removed from Ψ by the data cleaning process.
3) Therefore, if there are lots of tuples with ? or non-deterministic information in a table data set, the number of considerable tuples may become the small number of tuples. Most of tuples may be ignored by the data cleaning process.
4) The research on the data mining algorithm for handling tables like Ψ will be one solution for preserving personal information.

C. A Proposal of a Privacy-preserving Questionnaire by Non-deterministic Information

Proposal 1: We propose the following questionnaire *QUEST_Non-Det* by the organizer.

1) The questionnaire consists of some questions, which are multiple choices.
2) Each respondent may answer non-deterministic information for an inconvenient question. In this case, we may have tables like Ψ in Table II.
3) Each respondent intentionally preserves his personal information by using non-deterministic information.
4) *QUEST_Non-Det* may have vague information than *QUEST_Det*, so the organizer may not have the precise tendency of the respondents. However, the purpose of the questionnaire seems to know the overview of the respondents' tendency. Therefore, it will be useful to consider *QUEST_Non-Det* and its analysis.

However, we need new data mining algorithms for analyzing *QUEST_Non-Det*. In the subsequent section, we consider the analytical method for *QUEST_Non-Det*.

III. The Environment of Data Analysis for Questionnaire *QUEST_Non-Det*

This section follows the framework of Rough Non-deterministic Information Analysis (RNIA) [19] and apply it to analyzing Questionnaires *QUEST_Det* and *QUEST_Non-Det*.

A. Rules in DIS

This subsection considers rules in DIS, i.e., *QUEST_Det*. We usually fix an attribute Dec as the decision attribute, and handle a pair $[A, v]$ of the attribute A and its attribute value v, which we call a *descriptor*. An *implication* is a formula $\tau : \wedge_i [A_i, v_i] \Rightarrow [Dec, val]$, and we see an implication satisfying some constraints as a *rule*. In most of work on rule generation, the next two constraints are employed [15], [16], [26], [27], and we also employ them.

$$\begin{array}{l} \text{For two threshold values } 0 < \alpha, \beta \leq 1.0, \\ support(\tau)(= Num(\tau)/Number\_of\_the\_tuples) \geq \alpha, \\ accuracy(\tau)(= Num(\tau)/Num(\wedge_i [A_i, v_i])) \geq \beta. \\ Num(F) \text{ is the number of objects supporting a formula } F. \end{array} \quad (1)$$

Definition 1: For DIS with a decision attribute Dec, and threshold values α and β, an implication τ is a *rule*, if τ satisfies both $support(\tau) \geq \alpha$ and $accuracy(\tau) \geq \beta$.

object	attrib	value
r1	q1	1
r1	q2	1
r1	q3	1
r1	q4	1
r2	q1	1
r2	q2	2
r2	q3	1
r2	q4	1
r3	q1	2
r3	q2	2

Fig. 3. A part of a table Φ_{RDF} defined by Φ.

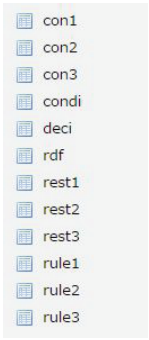


Fig. 4. The generated all tables.

Let us consider an implication $\tau : [q1, 1] \wedge [q3, 2] \Rightarrow [q4, 3]$ in Table I. Since τ is supported by the objects r_4 and r_5, we have $support(\tau)$=2/6=1/3, and the formula $[q1, 1] \wedge [q3, 2]$ is also supported by the objects r_4 and r_5. Therefore, $accuracy(\tau)$=2/2=1.0 holds. The $support(\tau)$ value means the ratio on the occurrence of τ, and the $accuracy(\tau)$ value means the ratio on the consistency of τ.

B. Apriori-based Rule Generation in DIS

This subsection describes rule generation in DIS toward rule generation in NIS.

1) An Implemented Software in SQL for Handling DISs: In RNIA, the Apriori algorithm for the transaction data [3], [4], [23] is adjusted to the algorithm for table data sets. Here, each item in the transaction data is identified with a descriptor $[A, v]$. We recently employed the environment phpMyAdmin [17], and implemented this Apriori algorithm in SQL procedures based on [19]. By using the actual execution of Φ in Table I, we describe the Apriori-based rule generation.

2) Data Sets in the RDF Format: We employ the RDF format [24], [25]. Figure 3 is a part of Φ_{RDF} in the RDF format of Φ. In the usual csv data sets, the attribute value val is assigned to the pair of the object r and the attribute $attrib$, namely we may see the csv data set is a set of all triplet $(r, attrib, value)$. Φ_{RDF} is a set of all triplet $(r, attrib, value)$, and each tuple in Φ_{RDF} corresponds to a descriptor. It is easy to generate a table data set in the RDF format from a csv file.

3) Rule Generation in Table Φ : We fix the decision attribute Dec, the threshold values α and β, then we execute the SQL query. Figure 5 shows the obtained all rules under the condition Dec='q4', α=0.3, and β=0.8.

This execution took about 1 (sec), and we had all tables in Figure 4. In the current program, it is possible to obtain the rules with less than three conditions. In the first step, tables $con1$ and $deci$ are generated, then the Cartesian product is generated, and finally the implications satisfying the constraints are stored in a table $rule1$. The implication τ satisfying $support(\tau) \geq \alpha$ and $accuracy(\tau) < \beta$ is stored in a table $rest1$. In the second step, a table $con2$ is generated

att1	val1	decision	support	accuracy	
q2	3	3	0.333	1.000	①
q3	1	1	0.333	1.000	②

att1	val1	att2	val2	decision	support	accuracy	
q1	1	q3	2	3	0.333	1.000	③

Fig. 5. The obtained all rules form Φ_{RDF}. In the above table, we have rules $[q2, 3] \Rightarrow [q4, 3]$, $[q3, 1] \Rightarrow [q4, 1]$, and $[q1, 1] \wedge [q3, 2] \Rightarrow [q4, 3]$.

from $rest1$, and the similar procedure is applied to $con2$ and $deci$. The same procedure is also applied to tables $con3$ and $deci$ in the third step.

Like this, it is possible to obtain a set of rules, which show us the tendency in Φ.

C. Rules in NIS

This section follows the framework of RNIA, and considers the rules in NIS, i.e., *QUEST_Non-Det*. In RNIA, the modal concepts, i.e., the certainty and the possibility are considered, and the certain rules and the possible rules are defined by using all tuples in NIS.

1) Derived DISs from NIS: In NIS, we replace each set in a table with a value in the set, and define one DIS, which is named a *derived* DIS from NIS. In Ψ, we have the 72 derived DISs, and we see an actual DIS is in the 72 derived DISs (Figure 6).

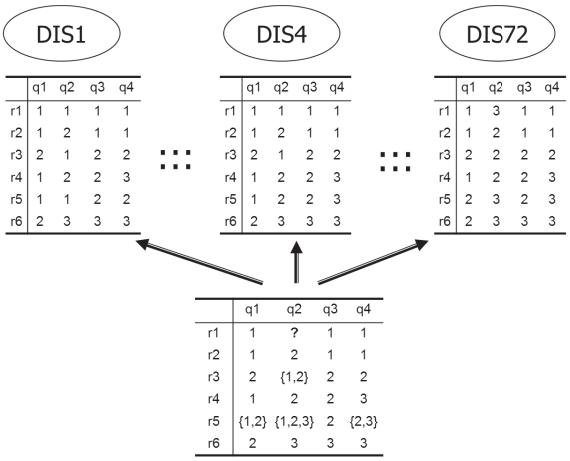
2) Certain Rules and Possible Rules in NIS: In RNIA, the following rules are defined based on all derived DISs, namely based on all tuples with non-deterministic information.

Definition 2: For NIS with a decision attribute Dec, and threshold values α and β, (1) and (2) are given.
(1) An implication τ is a *certain* rule, if τ is a rule in each derived DIS.
(2) An implication τ is a *possible* rule, if τ is a rule in at least one derived DIS.

The certain rule is a rule in the unknown actual derived DIS, and it is not influenced by information incompleteness. The possible rule may be a rule in the unknown actual derived DIS, and this rule is related to the possibility. There may be a case that both $[A, v] \Rightarrow [Dec, val_1]$ and $[A, v] \Rightarrow [Dec, val_2]$ are the possible rules at the same time.

These two types of rules seem to be the natural extension of rules in DIS, and we will be able to know the tendency of the respondents by using the certain rules and the possible rules.

3) A Problem for Handling the Certain and the Possible rules : However, we face with the problem that the number of derived DISs increases exponentially. Even in Ψ, the number of derived DISs is 72 (=$2^3 \times 3^2$). In Mammographic data set in the UCI machine leaning repository [8], the number of derived



	q1	q2	q3	q4
r1	1	1	1	1
r2	1	2	1	1
r3	2	1	2	2
r4	1	2	2	3
r5	1	1	2	2
r6	2	3	3	3

	q1	q2	q3	q4
r1	1	1	1	1
r2	1	2	1	1
r3	2	1	2	2
r4	1	2	2	3
r5	1	2	2	3
r6	2	3	3	3

	q1	q2	q3	q4
r1	1	3	1	1
r2	1	2	1	1
r3	2	2	2	2
r4	1	2	2	3
r5	2	3	2	3
r6	2	3	3	3

	q1	q2	q3	q4
r1	1	?	1	1
r2	1	2	1	1
r3	2	{1,2}	2	2
r4	1	2	2	3
r5	{1,2}	{1,2,3}	2	{2,3}
r6	2	3	3	3

Fig. 6. The 72 derived DISs from Ψ.

DISs is more than 10^{100}. Therefore, it will be hard to employ the typical method such that we sequentially pick up a derived DIS and examine the constraints.

4) Theoretical Properties in Rule Generation: We briefly follow the theoretical properties on rules. Let Ω and OB denote NIS and the set of objects (the set of respondents), respectively. Furthermore, let $DD(\Omega)$ denote the set of all derived DISs from Ω, and the following is defined in Ω.

$$\begin{array}{l}
(1)\ minsupp(\tau) = min_{\omega \in DD(\Omega)}\{support(\tau) \text{ in } \omega\},\\
(2)\ minacc(\tau) = min_{\omega \in DD(\Omega)}\{accuracy(\tau) \text{ in } \omega\},\\
(3)\ maxsupp(\tau) = max_{\omega \in DD(\Omega)}\{support(\tau) \text{ in } \omega\},\\
(4)\ maxacc(\tau) = max_{\omega \in DD(\Omega)}\{accuracy(\tau) \text{ in } \omega\},\\
\text{If } \tau \text{ does not occur in } \omega, \text{ we deine}\\
support(\tau) = accuracy(\tau) = 0 \text{ in } \omega.
\end{array} \tag{2}$$

The above definitions depend upon each $\omega \in DD(\Omega)$, however it is possible to calculate these four values based on [19]. Furthermore, this calculation does not depend upon the number of $DD(\Omega)$.

For calculating the above values, at first the following two sets of objects, i.e., *inf* and *sup* blocks, are defined for each descriptor $[A_i, v_i]$.

$$\begin{array}{l}
(1)\ inf([A_i, v_i]) = \{r \in OB \mid \text{the attribute value is deterministic and it is } v_i\},\\
(2)\ inf(\wedge_i[A_i, v_i]) = \cap_i\ inf([A_i, v_i]),\\
(3)\ sup([A_i, v_i]) = \{r \in OB \mid \text{the attribute value is non-deterministic and } v_i \text{ is in the set}\},\\
(4)\ sup(\wedge_i[A_i, v_i]) = \cap_i\ sup([A_i, v_i]).
\end{array} \tag{3}$$

Based on the definitions of inf and sup blocks, the following calculation formula for $minsupp(\tau)$ to $maxacc(\tau)$ are given. For *NIS* Ω, let τ be an implication $\wedge_i[A_i, v_i] \Rightarrow [Dec, val]$. Then, the following is shown in [19].

(1) For τ which occurs in each $\omega \in DD(\Omega)$,
$minsupp(\tau)$
$= |inf(\wedge_i[A_i, v_i]) \cap inf([Dec, val])|/|OB|$,
$minacc(\tau)$
$= \frac{|inf(\wedge_i[A_i,v_i]) \cap inf([Dec,val])|}{|inf(\wedge_i[A_i,v_i])|+|OUT|}$.
Otherwise,
$minsupp(\tau) = minacc(\tau) = 0$.

(3) $maxsupp(\tau)$
$= |sup(\wedge_i[A_i, v_i]) \cap sup([Dec, val])|/|OB|$. (4)

(4) $maxacc(\tau)$
$= \frac{|inf(\wedge_i[A_i,v_i]) \cap sup([Dec,val])|+|IN|}{|inf(\wedge_i[A_i,v_i])|+|IN|}$.

(5) OUT
$= \{sup(\wedge_i[A_i, v_i]) \setminus inf(\wedge_i[A_i, v_i])\}$
$\setminus inf([Dec, val])$.

(6) IN
$= \{sup(\wedge_i[A_i, v_i]) \setminus inf(\wedge_i[A_i, v_i])\}$
$\cap sup([Dec, val])$.

Furthermore, the next properties are shown in [19].

1) There exists at least one $\omega_{min} \in DD(\Omega)$ which makes both values of $support(\tau)$ and $accuracy(\tau)$ the minimum.
2) There exists at least one $\omega_{max} \in DD(\Omega)$ which makes both values of $support(\tau)$ and $accuracy(\tau)$ the maximum.

Namely, Figure 7 holds for each τ, and Theorem 1 is concluded.

Theorem 1: The following holds for NIS.
(1) τ is a certain rule, if and only if there is τ satisfying $minsupp(\tau) \geq \alpha$ and $minacc(\tau) \geq \beta$ (Figure 8).
(2) τ is a possible rule, if and only if there is τ satisfying $maxsupp(\tau) \geq \alpha$ and $maxacc(\tau) \geq \beta$ (Figure 9).
(3) Both conditions in (1) and (2) do not depend upon the number of $DD(\Omega)$.

Example 1: Let us apply the properties and Theorem 1 to Ψ under the same condition in Figure 5.
(1) For $\tau : [q3, 1] \Rightarrow [q4, 1]$, the following holds.
$inf([q3, 1])$=$sup([q3, 1])$=$\{r1, r2\}$,
$inf([q4, 1])$=$sup([q4, 1])$ =$\{r1, r2\}$,
OUT=$(sup([q3, 1]) \setminus inf([q3, 1])) \setminus inf([q4, 1])$=$\emptyset$,
IN=$(sup([q3, 1]) \setminus inf([q3, 1])) \cap sup([q4, 1])$=$\emptyset$,
$minsupp(\tau)$=$|inf([q3, 1]) \cap inf([q4, 1])|/|OB|$
=$|\{r1, r2\}|$/6=1/3,
$minacc(\tau)$=$|\{r1, r2\}|/(|inf([q3, 1])| + |OUT|)$
=$|\{r1, r2\}|/(|\{r1, r2\}| + |\emptyset|)$=2/2=1.0.
Since $minsupp(\tau) \geq 0.3$ and $minacc(\tau) \geq 0.8$ hold, we conclude the implication τ is a certain rule by using Figure 8. Namely, this τ is always a rule in each $\omega \in DD(\Psi)$. Clearly, the above procedure does not depend upon the number of $DD(\Psi)$.
(2) For $\tau : [q1, 1] \wedge [q3, 2] \Rightarrow [q4, 3]$, the following holds.

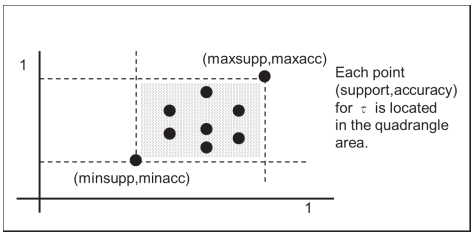


Fig. 7. Each point $(support(\tau), accuracy(\tau))$ by ω.

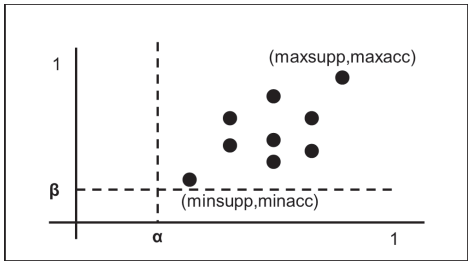


Fig. 8. The case of examining the certain rule.

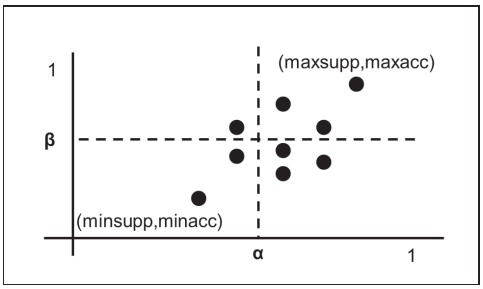


Fig. 9. The case of examining the possible rule.

$inf([q1, 1]) \cap inf([q3, 2])$=$\{r1, r2, r4\} \cap \{r3, r4, r5\}$=$\{r4\}$,
$sup([q1, 1]) \cap sup([q3, 2])$=$\{r1, r2, r4, r5\} \cap \{r3, r4, r5\}$
=$\{r4, r5\}$,
$inf([q4, 3])$=$\{r4, r6\}$, $sup([q4, 3])$=$\{r4, r5, r6\}$,
OUT=$(\{r4, r5\} \setminus \{r4\}) \setminus \{r4, r6\}$=$\{r5\}$,
IN=$(\{r5\}) \cap \{r4, r5, r6\}$=$\{r5\}$,
$minsupp(\tau)$=$|\{r4\} \cap inf([q4, 3])|/|OB|$=$|\{r4\}|$/6=1/6,
$minacc(\tau)$=$|\{r4\}|/(|\{r4\}| + |\{r5\}|)$=0.5,
$maxsupp(\tau)$=$|\{r4, r5\} \cap sup([q4, 3])|/|OB|$=$|\{r4, r5\}|$/6=1/3,
$maxacc(\tau)$=$(|\{r4\} \cap \{r4, r5, r6\}| + |IN|)/(|\{r4\}| + |IN|)$
=$(|\{r4\}| + |\{r5\}|)/(|\{r4\}| + |\{r5\}|)$=1.0.
Since $minsupp(\tau) < 0.3$ and $minacc(\tau) < 0.8$ hold, we conclude the implication τ is not any certain rule, but $maxsupp(\tau) \geq 0.3$ and $maxacc(\tau) \geq 0.8$ hold. Therefore, we conclude τ is a possible rule.

D. Apriori-based Rule Generation in NIS

In RNIA, two rule generation systems are given, namely the certain rule generator and the possible rule generator. Both systems extend the Apriori algorithm in DIS to NIS by using Theorem 1. We describe Apriori-based rule generation by using the actual execution of Ψ.

1) An Overview of NIS-Apriori based Rule Generation: In the certain rule generator, tables $con1$ (the condition part) and $deci$ (the decision part) in Figure 4 are generated at first, then for each τ, $minsupp(\tau)$ and $minacc(\tau)$ are calculated. By using Theorem 1, the certain rule generator decides τ is a certain rule or not. If $minsupp(\tau) \geq \alpha$ and $minacc(\tau) \geq \beta$, this τ is stored in a table $crule1$. If $minsupp(\tau) \geq \alpha$ and $minacc(\tau) < \beta$, this τ is stored in a table $crest1$. By using $crest1$, the next $ccon2$ is generated.

In the possible rule generator, $maxsupp(\tau)$ and $maxacc(\tau)$ are calculated by using tables $con1$ and $deci$. By using Theorem 1, the possible rule generator decides τ is a possible rule or not. If $maxsupp(\tau) \geq \alpha$ and $maxacc(\tau) \geq \beta$, this τ is stored in a table $prule1$. If $maxsupp(\tau) \geq \alpha$ and $maxacc(\tau) < \beta$, this τ is stored in a table $prest1$. By using $prest1$, a table $pcon2$ is generated.

Like this, rule generation in DIS is extended to rule generation in NIS Ω. Since every calculation of $minsupp(\tau)$, $minacc(\tau)$, $maxsupp(\tau)$, and $maxacc(\tau)$ does not depend upon $DD(\Omega)$ and each calculation is the polynomial time order, the extended algorithm will take about the twice complexity of the Apriori algorithm in DIS. In RNIA, this algorithm is called the *NIS-Apriori* algorithm [19].

2) An Implemented Software in SQL for Handling NISs: We also employed the environment phpMyAdmin [17], and implemented this NIS-Apriori algorithm in SQL procedures based on [19]. By using the actual execution of Ψ, we describe Apriori-based rule generation in NISs.

3) Data Sets in the RDF Format: We employ a table Ψ_{NRDF} in the NRDF format for Ψ (Figures 10 and 11). In Ψ_{NRDF}, each tuple corresponds to a descriptor in one derived DIS. The fourth attribute det=1 means the tuple is definite. Otherwise, each tuple comes from non-deterministic information. It is also easy to generate a table data set in the NRDF format.

4) Rule Generation in Table Ψ: We employ the same conditions in Φ, namely we fix the condition Dec='q4', α=0.3, and β=0.8. In NIS, we prepared three procedures step1, step2, and step3 in SQL. The step1 generates rules with one condition, the step2 does rules with two conditions, and the step3 does rules with three conditions.

It took about 2 (sec) for executing each step1, step2, and step3. We obtained all tables in Figure 12. Figure 13 and 14 show the obtained all certain rules and all possible rules.

Here, we clarify the relation between rules in DIS and rules in NIS, and we have the following remarks.

Remark 1: For NIS Ω, let ω be a derived DISs from Ω. Then, any rule in ω is at least a possible rule in Ω.

Remark 2: The property about *QUEST_Det* and *QUEST_Non-Det*

1) Some rules in *QUEST_Det* may be obtained as certain rules in *QUEST_Non-Det*, for example, ② in Figure 5 and Figure 13.

object	attrib	value	det
r1	q1	1	1
r1	q2	1	3
r1	q2	2	3
r1	q2	3	3
r1	q3	1	1
r1	q4	1	1
r2	q1	1	1
r2	q2	2	1
r2	q3	1	1
r2	q4	1	1

Fig. 10. A part of a table Ψ_{NRDF} defined by Ψ.

r3	q1	2	1
r3	q2	1	2
r3	q2	2	2
r3	q3	2	1
r3	q4	2	1
r4	q1	1	1
r4	q2	2	1
r4	q3	2	1
r4	q4	3	1
r5	q1	1	2

Fig. 11. A part of a table Ψ_{NRDF} defined by Ψ.

Fig. 12. The obtained all tables.

att1	val1	deci	deci_value	minsupp	minacc	
q3	1	q4	1	0.333	1.000	②

Fig. 13. The obtained one certain rule $[q3, 1] \Rightarrow [q4, 1]$.

att1	val1	deci	deci_value	maxsupp	maxacc	
q2	1	q4	2	0.333	1.000	possible
q2	3	q4	3	0.333	1.000	①
q3	1	q4	1	0.333	1.000	possible

a1	v1	a2	v2	deci	deci_value	maxsupp	maxacc	
q1	1	q3	2	q4	3	0.333	1.000	③
q1	2	q2	2	q4	2	0.333	1.000	possible
q1	2	q3	2	q4	2	0.333	1.000	possible
q2	2	q3	2	q4	3	0.333	1.000	possible

Fig. 14. The obtained all possible rules.

2) Any rule in *QUEST_Det* is at least obtained as a possible rule in *QUEST_Non-Det*, for example, ① and ③ in Figure 5 and Figure 14.
3) Other independent rules may be obtained as possible rules like the rules marked 'possible' in Figure 14.

IV. Discussion about QUEST_Det and QUEST_Non-Det

This section enumerates the merit and the demerit of *QUEST_Det* and *QUEST_Non-Det* by using Figure 15. We also discuss about the use of the privacy-preserving questionnaire by non-deterministic information.

Remark 3: A case of *QUEST_Det* Φ

Merit:

The precise information is stored in Φ ((i) in Figure 15), so it is possible to obtain the correct rules and the tendency of the respondents ((i) and (ii), Figure 15).

Demerit:

Since the precise information is stored in Φ ((i) in Figure 15), it is necessary to pay attention to manage Φ. Recently, we often have the leaks of data sets. The privacy-preserving functionality is required ((ii) in Figure 15).

Remark 4: A case of *QUEST_Non-Det* Ω

Merit:

Information is diluted from Φ to Ω ((iii) in Figure 15), which defines $DD(\Omega)$. This Ω is more privacy-preserved than Φ. This Ω may be effective for the leaks of the data sets. Instead of the rules in Φ, the certain rules and the possible rules are obtained in Ω. These two rules are defined by all tuples with non-deterministic information, and we have the weakened results from Φ, but the privacy in Ω is more preserved than that in Φ.

Demerit:

The certain rule is a rule in Φ, but there may be several possible rules independent from Φ. In RNIA, we cannot discriminate the possible rules in Φ with the possible rules independent from Φ.

Remark 5: The trade-off relation between Φ and Ω

1) We have the correct results from the precise information, i.e., Φ, and we have the weakened results from the vague information, i.e., Ω.
2) On the other hand, information in Ω is more privacy-preserved than that in Φ.
3) Probably, this types of questionnaire has not been considered, because many tuples may be ignored based on the information cleaning process. However, the framework of RNIA gives a new possibility for the privacy-preserving questionnaire.
4) In (iv), Figure 15, there seems less software tools except the software in RNIA.

In this subsection, we clarified the merit and the demerit of *QUEST_Det* and *QUEST_Non-Det*, and the possibility for the privacy-preserving questionnaire.

V. Two Examples by NIS-Apriori in SQL

This section gives two examples by the software NIS-Apriori in SQL. The one is data mining from Mammographic data set in the UCI machine learning repository [8], and the other is the questionnaire based on Lenses data set in this repository.

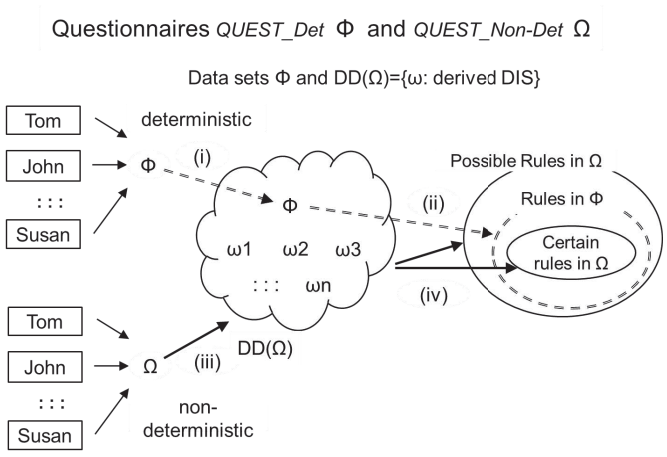


Fig. 15. The survey of *QUEST_Det* and *QUEST_Non-Det* .

A. A Case of Mammographic Data Set

Mammographic data set, which handles the cancer data, consists of 960 objects, 6 attributes ($assessment$, age, $shape$, $margin$, $density$, and $severity$). We can see the $shape$ and $density$ as four-choice questions, the $margin$ as five-choice question, and the decision $severity$ as two-choice (1:benign and 2:malignant) question.

There are about 150 missing values, and we replace each missing value to a set of all choices, and we obtained NIS Σ. If we employ *QUEST_Non-Det*, we will have the similar data set as Σ. In this case, the number of derived DISs is more than 10^{100}, so it will be hard to obtain the certain rules and the possible rules without Theorem 1. Figure 16 and 17 show the obtained rules from Σ. It took about 20(sec) for $step1$ and 7(sec) for $step2$, and we obtained rules with only one condition. Since ④ to ⑦ are certain rules, we conclude that they also hold in the unknown actual derived DIS.

B. A Case of the Lenses Data Set

Lenses data set, which handles the contact lenses data, consists of 24 objects, five attributes age with three choices 1, 2, 3, $spec$, $asti$, $tear$ with two choices 1, 2, and the decision attribute dec with three choices 1: hard contact lenses, 2: soft contact lenses, 3: no lenses.

In this data set, there is no missing values, and we see this data set as one DIS π, namely we see π is the obtained data set by *QUEST_Det*. We randomly added non-deterministic information to π, and generated one NIS Π, which we see the obtained data set by *QUEST_Non-Det*. In Π, the 30 attribute values in π are changed to non-deterministic information. This means 25% (=30 /120) information of π is hidden, and $DD(\Pi)$ consists of 8153726976 (=$2^{25} \times 3^5$) derived DISs. The DIS π is an element of $DD(\Pi)$. Figure 18 shows the obtained rules in π, and Figure 19 shows the obtained all possible rules in

att1	val1	deci	deci_value	minsupp	minacc	
assess	4	severity	0	0.445	0.763	④
assess	5	severity	1	0.317	0.869	⑤
margin	1	severity	0	0.329	0.859	⑥
shape	4	severity	1	0.328	0.752	⑦

Fig. 16. The obtained all certain rules from Σ under the condition Dec=severity, α=0.2, and β=0.7.

att1	val1	deci	deci_value	maxsupp	maxacc	
assess	4	severity	0	0.451	0.783	④
assess	5	severity	1	0.330	0.888	⑤
margin	1	severity	0	0.368	0.896	⑥
shape	1	severity	0	0.214	0.844	possible
shape	2	severity	0	0.203	0.848	possible
shape	4	severity	1	0.341	0.794	⑦

Fig. 17. The obtained all possible rules from Σ under the condition Dec=severity, α=0.2, and β=0.7.

att1	val1	decision	support	accuracy	
astigmatic	2	3	0.333	0.667	⑧
spec	2	3	0.333	0.667	⑨
tear	1	3	0.500	1.000	⑩

Fig. 18. The obtained all rules from π under the condition Dec=decision, α=0.3, and β=0.6.

att1	val1	deci	deci_value	maxsupp	maxacc	
age	1	decision	3	0.333	0.667	
age	2	decision	3	0.333	0.727	
astigmatic	1	decision	3	0.458	0.688	
astigmatic	2	decision	3	0.333	0.667	⑧
spec	1	decision	3	0.458	0.688	
spec	2	decision	3	0.458	0.733	⑨
tear	1	decision	3	0.542	1.000	⑩
tear	2	decision	3	0.583	0.609	

Fig. 19. The obtained all possible rules from Π under the condition Dec=decision, α=0.3, and β=0.6.

Π. It took about 1 (sec) for generating rules in π, and it took about 2 (sec) in $step1$ for Π. We did not obtain any certain rules.

Like this, we will be able to consider rules not only in *QUEST_Det* but also in *QUEST_Non-Det*.

VI. Some Examples by *getRNIA* System

We have also implemented a software tool *getRNIA* (`http://getrnia.org`) [28] in Python based on NIS-Apriori algorithm. This is a demonstrative software tool for showing rule generation in NIS, and we can easily execute some demo files, for example, Soybean.csv, Congress.csv, Cancer.csv, Mammographic.csv, Hepatitis.csv, etc. in Figure 20. They are all picked up from the UCI machine learning repository.

Fig. 20. Demo files in *getRNIA*.

Figure 21 shows the revised Congress.csv data set, which consists of the 435 objects, the 17 attributes, the 392 missing values, and the number of derived DISs is more than 10^{100}. In the first attribute, the set of the attribute values is $\{democrat, republican\}$, and the other set of the attribute values is $\{yes, no\}$. Each missing value '?' is changed to a set of all possible values, i.e., $\{yes, no\}$.

Figure 22 shows the screen shot on rule generation, where the constraints are $support(\tau) \geq 0.5$ and $accuracy(\tau) \geq 0.7$. The blue circle implies the minimum point ($minsupp(\tau)$,$minacc(\tau)$), and the orange circle implies the maximum point ($maxsupp(\tau)$,$maxacc(\tau)$). Even though there are more than 10^{100} derived DISs in Congress data set, each point is located in the rectangle area defined by the minimum point and the maximum point. By checking these two characteristic points, we can obtain rules depending upon more than 10^{100} derived DISs. In *getRNIA*, we can easily change the constraints, and we obtain rules defined by new constraints. However, *getRNIA* handles rules in the form of either $Con_1 \Rightarrow Decision$ or $Con_1 \wedge Con_2 \Rightarrow Decision$ for reducing the execution time.

VII. Concluding Remarks

We followed the framework of RNIA, and proposed new questionnaire by using non-deterministic information. In this questionnaire, each respondent intentionally preserves his privacy, and we can analyze such data sets. Based on [19], we implemented the prototype of NIS-Apriori in SQL, and applied this software to some data sets like Mammographic data set and Lenses data set. Furthermore, we showed the screen shot

of the execution by *getRNIA*. In our proposal, the background is rule generation in NIS, namely NIS-Apriori based rule generator. Based on the execution logs by the prototype in SQL and *getRNIA*, we think that the background is proved to be robust and stable.

Toward the actual application, we need to consider the details below.

(1) About non-deterministic information and the missing values: The use of non-deterministic information and the missing values will be effective for privacy-preserving. However in the actual questionnaire, we will employ binary choices, i.e., 'either A or B', in the multi-choice question. Namely, we will handle the answers with 2-anonymity. Even though we can replace a missing value with non-deterministic information 'either A_1, A_2, $\cdots$, or A_n', we should escape from this choice, because such choice causes the more weakened results.

(2) About the software: We are coping with the software tool in SQL, because SQL has the high versatility. We simply simulated the NIS-Apriori algorithm by using the procedures in SQL, and realized the next procedures [20], [21],

(i) $apri(Dec, |OB|, \alpha, \beta)$ for RDF format,
(ii) $step1(Dec, |OB|, \alpha, \beta)$ for NRDF format,
(iii) $step2(Dec, |OB|, \alpha, \beta)$ for NRDF format,
(iv) $step3(Dec, |OB|, \alpha, \beta)$ for NRDF format.

In NRDF format, if we consider a case that every det=1, this corresponds to RDF format. We checked some data sets, and examined that each procedure generated the same tables. In the current implementation, we faithfully simulated the NIS-Apriori algorithm, so we had several temporal tables. The current procedures are just the prototype, and we need to refine the procedures.

(3) About the actual questionnaire data analysis: In this paper, we employed several NISs instead of the actual questionnaire, because the proposing questionnaire defines one NIS. Rule generation in NIS is the main issue in the proposing questionnaire. In the next research, we need to handle the actual questionnaire, and it is necessary to consider the analysis depending upon the property of a questionnaire.

ACKNOWLEDGMENT

The first author would be grateful to Prof. Dominik Ślęzak for his guidance to SQL. This work is supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Number 26330277.

REFERENCES

[1] Aggarwal, C., Yu, S. (Eds): Privacy-preserving data mining: models and algorithms. Advances in Database Systems 34, Springer (2008)
[2] Aggarwal, C., Yu, S.: A general survey of privacy-preserving data mining models and algorithms. Advances in Database Systems 34, Springer (2008)
[3] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. Proc. VLDB'94 Morgan Kaufmann, 487–499 (1994)
[4] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining AAAI/MIT Press, 307–328 (1996)
[5] Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. Proc. ACM SIGMOD Conference, 439-450 (2000)
[6] Bogdanov, D., Talviste, R.: A prototype of online privacy-preserving questionnaire system. Technical report, University of Tartu (2010)
[7] Clark, P., Grzymala-Busse, J.: An analysis of probabilistic approximations for rule induction from incomplete data sets. Fundamenta Informaticae 132(3), 365-379 (2014)
[8] Frank, A., Asuncion, A.: UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science (2010) http://mlearn.ics.uci.edu/MLRepository.html
[9] Kikuchi, H.: Privacy preserving data mining. Proc. FIT2004, (in Japanese) (2004)
[10] Lipski, W.: On semantic issues connected with incomplete information databases. ACM Transactions on Database Systems 4(3), 262–296 (1979)
[11] Lipski, W.: On databases with incomplete information. Journal of the ACM 28(1), 41–70 (1981)
[12] Nakata, M., Sakai, H.: Twofold rough approximations under incomplete information. Int'l. J. General Systems 42(6), 546–571 (2013)
[13] Nakazato, J., Fujimoto, K., Kikuchi, H.: Privacy preserving web-based questionnaire. Proc. IEEE 19th Int'l. Conf. on Advanced Information Networking and Applications, 285–288 (2005)
[14] Orłowska, E., Pawlak, Z.: Representation of nondeterministic information. Theoretical Computer Science 29(1-2), 27–39 (1984)
[15] Pawlak, Z.: Systemy Informacyjne: Podstawy Teoretyczne (in Polish) WNT (1983)
[16] Pawlak, Z.: Rough Sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers (1991)
[17] phpMyAdmin Web Page http://www.phpmyadmin.net/ (2016)
[18] Sakai, H., Ishibashi, R., Koba, K., Nakata, M.: Rules and apriori algorithm in non-deterministic information systems. Transactions on Rough Sets 9, 328–350 (2008)
[19] Sakai, H., Wu, M., Nakata, M.: Apriori-based rule generation in incomplete information databases and non-deterministic information systems. Fundamenta Informaticae 130(3), 343–376 (2014)
[20] Sakai, H., Liu, C., Zhu X., Nakata, M.: On NIS-Apriori based data mining in SQL. Proc. Int'l. Conf. on Rough Sets 2016, Springer LNAI 9920, 514–524 (2016)
[21] Sakai, H.: Software Tools for RNIA (Rough Non-deterministic Information Analysis) Web Page (2016) http://www.mns.kyutech.ac.jp/~sakai/RNIA/
[22] Sakuma, J., Kobayashi, S.: Privacy-preserving data mining. Journal of the Japanese Society of AI 26(5), 1–11 (2011)
[23] Sarawagi, S., Thomas, S., Agrawal, R.: Integrating association rule mining with relational database systems: alternatives and implications. Data Mining and Knowledge Discovery 4(2), 89–125 (2000)
[24] Ślęzak, D., Sakai, H.: Automatic extraction of decision rules from non-deterministic data systems: Theoretical foundations and SQL-based implementation. DTA2009 Springer CCIS Vol.64, 151–162 (2009)
[25] Swieboda, W., Nguyen, S.: Rough set methods for large and spare data in EAV format. Proc. IEEE RIVF 2012, 1–6 (2012)
[26] Tsumoto, S.: Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Information Sciences 124(1-4), 125–137 (2000)
[27] Tsumoto, S.: Automated extraction of hierarchical decision rules from clinical databases using rough set model. Expert Systems with Applications 24, 189–197 (2003)
[28] Wu, M., Nakata, M., Sakai, H.: An overview of the getRNIA system for non-deterministic data. Procedia Computer Science 22, 615–622 (2013) http://getrnia.org

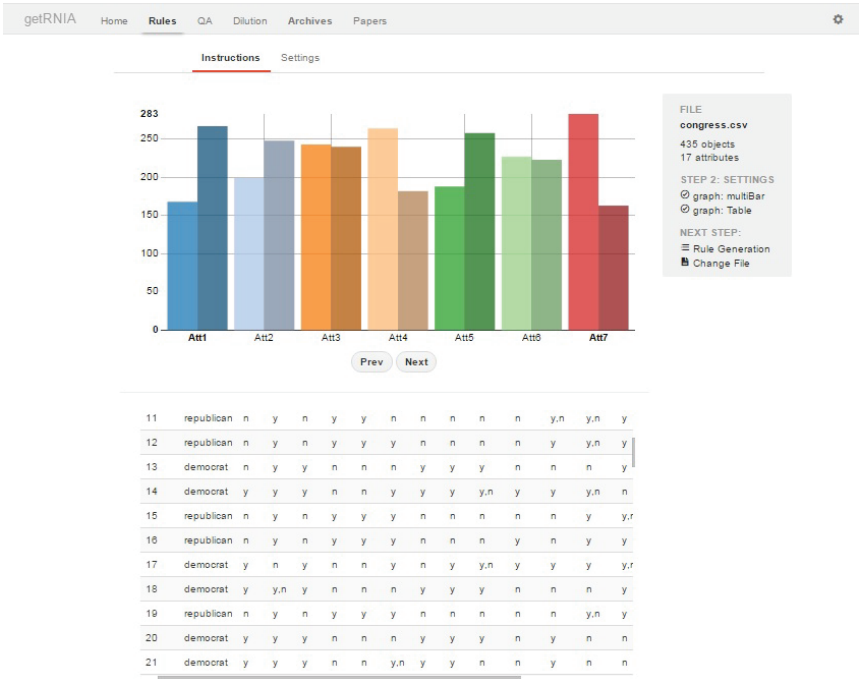
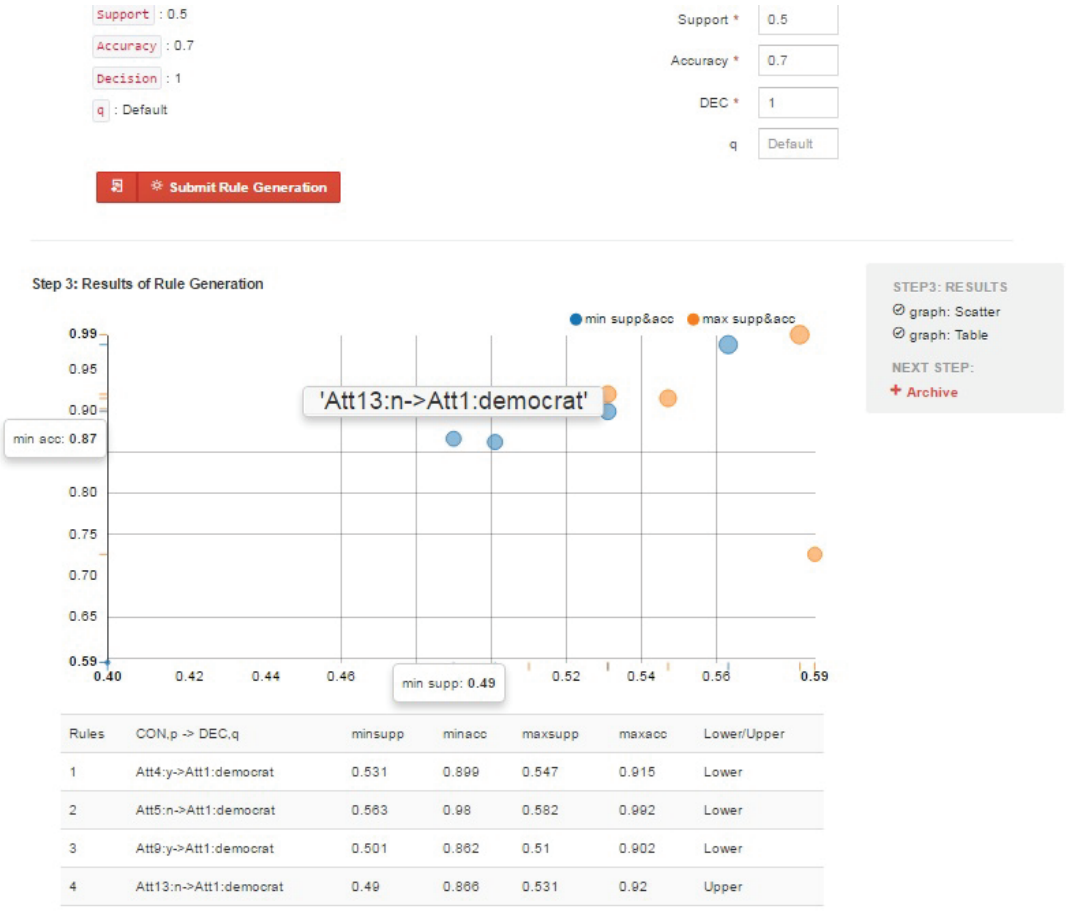


Fig. 21. Congress data set, where each missing value is changed to the possible values {yes,no}. There are more than 10^{100} derived DISs.



Rules	CON,p -> DEC,q	minsupp	minacc	maxsupp	maxacc	Lower/Upper
1	Att4:y->Att1:democrat	0.531	0.899	0.547	0.915	Lower
2	Att5:n->Att1:democrat	0.563	0.98	0.582	0.992	Lower
3	Att9:y->Att1:democrat	0.501	0.862	0.51	0.902	Lower
4	Att13:n->Att1:democrat	0.49	0.866	0.531	0.92	Upper

Fig. 22. The generated certain rules (Lower) and the possible rules (Upper). The implication $[Att13, n] \Rightarrow [Att1, democrat]$ does not satisfy the constraints for a certain rule, but it satisfies the constraints for a possible rule. An implication $[Att4, y] \Rightarrow [Att1, democrat]$ satisfies the constraints in each of more than 10^{100} derived DISs.